LA-UR-21-31709
Approved for public release; distribution is unlimited.

Title: Real-time data reduction at 100 Tbps: Challenge and opportunity for AI-based data reduction for next-generation large-scale nuclear physics collider experiment

Author(s): Liu, Ming Xiong
Huang, Jin
Miryala, Sandeep
Ren, Yihui

Intended for: AI at DOE-Round Table Discussion

Issued: 2021-11-30

Real-time data reduction at 100 Tbps: Challenge and opportunity for AI-based data reduction for next-generation large-scale nuclear physics collider experiment

Jin Huang[1], Ming Liu[2], Sandeep Miryala[3], Yihui Ren[4]
1: Physics Department, Brookhaven National Laboratory
2: Physics Division, Los Alamos National Laboratory
3: Instrumentation Division, Brookhaven National Laboratory
4: Computational Science Initiative, Brookhaven National Laboratory
jhuang@bnl.gov, mliu@lanl.gov, smiryala@bnl.gov and yren@bnl.gov

Thrust: Impact of AI Advancements

Challenge

The modern large-scale nuclear physics (NP) experiments in high-energy particle colliders utilize streaming-readout electronics to digitize detector response at $O(100)$ **Tbps bandwidth**. Prominent examples at Brookhaven National Lab (BNL) include the sPHENIX experiment at Relativistic Heavy Ion Collider (RHIC) [4], which is under construction, and the experiments proposed for the Electron-Ion Collider (EIC) [5], planned for the 2030s . One of the main challenges for these streaming readout systems is to manage the data rate with **sufficient data reduction in real time** so the end-data fit persistent storage for offline analysis, which is typically at $O(1000)$ times smaller and $O(100)$ Gbps. Such data reduction traditionally is achieved via real-time high level triggers, which select and save a small subset of collisions of interest. Although triggering is applicable to high energy collider experiments such as those at the Large Hardron Collider at CERN, it is insufficient for these nuclear physics experiments which study diverse collision topologies. And traditional triggering approach is inefficient to preserve the max information harvested from the operation of colliders that costs $O(100)$M per year to DOE. Meanwhile, in recent years, ML-based high-throughput data reduction has emerged as a promising approach to efficiently preserve max information for a given space of persistent storage, e.g. via AI data compression, feature extraction, and noise filtering [1, 3, 2].

Collider experiments have widely used lossless real-time compression with a moderate compression ratio (~ 2). Lossy compression has better performance but requires longer development/validation cycles and may necessitate highly specialized hardware (e.g., ASICs) and a dedicated computing facility (e.g.ALICE O^2). High Energy Physics(HEP) / Nuclear Physics (NP) collider experiments, including sPHENIX, generate three-dimensional (3D) sparse data, presenting a unique challenge for lossy compression algorithms. For example, the data entries are encoded using 10-bit integers, whereas the current generic scientific lossy compression algorithms, such as SZ, ZFP, and MGARD are mostly designed for dense data and high-precision floating point entries, resulting sub-optimal compression rates and reconstruction fidelity. Such large volumes of sparse integer entries present a unique challenge for developing lossy compression algorithms.

Opportunity

Modern deep neural network (DNN) techniques and AI-centric hardware innovations present perfect opportunities to revolutionize real-time data processing pipelines in large HEP/NP experiments. In particular, DNN-based data compression, co-design and utilization of cutting-edge hardware. Compressing high-dimension data into low dimension embedding space has been a focal research topic in coding theory. Both lossy and lossless variations of compression algorithms have broad applications for audio, image and scientific data. However, existing methods have failed to meet the large-volume real-time processing throughput constraints: either too slow or too lossy. DNN approach provides a unique angle to tackle data compression problems.

First, an algorithm or trained DNN model is tailored to the characteristics of the data set, unlike general compression algorithms either by compressing unused bits or heuristics such as linearity of data. The manifold hypothesis states that for naturally occurred high-dimension data there exists a transform to map the data onto a low-dimension manifold. DNN models, trained through back-propagation, are the easiest way to achieve an approximation of such mapping by far. However, the quality of DNN models, besides the design of model architecture, heavily relies on the quality of training data and quality verification metrics. Comparing to traditional process of modeling (manually finding patterns in data and crafting heuristics to patch existing models), developing DNN models for scientific applications hinged on creating more realistic

and diverse training data, articulating non-trivial verification metrics to safeguard the model, and seeking ways to integrate known knowledge such as conservation laws into the DNN models.

Second, DNN models, consisting layers of linear transformations such as linear layers and convolutional layers followed by non-linear activations such as rectified linear units and hyperbolic tangent units, are extremely easy to compute in parallel on massively parallel processors such as graphic processing units (GPUs). Due to AI's success and its market potential, growing number of chip vendors and start up companies have optimized their products towards DNN training and inference. Innovative companies such as GraphCore, SambaNova, Cerebras and Xilinx have developed products based on the dataflow architecture. Deviating from the traditional von Neumann architecture where data are stored at a central memory and pulled (and pushed) through cache hierarchy, in a dataflow architecture, to reduce energy consumption and data fetching latency, memory and process units are divided into inter-connected tiles. This rapidly development of newer architecture potentially provides higher throughput and lower latency for real-time DNN model inference.

Timeliness and Research Directions

The sPHENIX experiment is coming online in 2023 [4]. And the EIC received DOE Mission Need status (CD-0) status in 2020 and is progressing rapidly through conceptual experimental design [5]. Therefore, it is an opportune moment to carry out this research, which has immediate benefits in the sPHENIX experiment and may radically impact the design of the future experiments in EIC, the next major NP facility funded by DOE. Preliminary results suggest that neural compression is superior both in compression ratio and reconstruction fidelity than traditional methods for sparse integer-valued TPC data [1]. For the next 2 to 3 years, there are three immediate research directions. First of all, neural compression algorithms can be extended to handle the entire accelerator detector input as a whole rather than at the sectional ($O(10^6)$ entries) level or 1/24 of the detector. Intuitively, we expect such model would improve the compression ratio further as inter-sectional relations can also be learned and compressed. Secondly, unlike most of the conventional compression methods where sizes of compressed data depend on the input data such as the number of zeros or entropy, compression ratio of neural compression is tuned and then fixed during the network architecture design process. Having an autonomous way to optimize the network architecture or varying the size of compressed data dynamically would be good to have. Thirdly, characterizing the throughput and latency performance of neural compression algorithm on different accelerators would inform the feasibility and deployment of such algorithms. As for the long-term, with the growing landscape of novel hardware, ever-improving DNN models and expending applications on experiment settings, a continuous integration system that can quickly 1) optimize and test existing models on novel hardware, 2) accommodate new datasets and verification methods, and 3) health check the real-time model performance and estimate model boundaries.

The impacts of a successful deployment of the AI real-time data reduction at sPHENIX and EIC are immediate and significant: an AI-driven reduction system would enable accessing almost 100% of all collisions. For the future EIC, it is the consensus of the community to have a full streaming DAQ to achieve the ultimate physics goals, in parallel with a conventional trigger system. An AI-driven systems are promising to deliver the required reduction performance optimizing the physics performance and hardware cost. We would also expect the technique and experience in these experiments to impact other fields at DOE complex where a high throughput data stream is desired, such as synchrotron light sources and electron microscopy imaging.

[1] Y. Huang et al. "Efficient Data Compression for 3D Sparse TPC via Bicephalous Convolutional Autoencoder". In: *IEEE 2021 International Conference on Machine Learning and Applications*. IEEE ICMLA". 2021. arXiv: 2111.05423 [cs.LG].

[2] *Intelligent Experiments Through Real-Time AI: Fast Data Processing and Autonomous Detector Control for sPHENIX and Future EIC Detectors*. DOE-FOA-0002490. 2021.

[3] S. Miryala et al. "Waveform Processing Using Neural Network Algorithms on the Front-end Electronics". In: *22nd International Workshop on Radiation Imaging Detectors*. 2021.

[4] sPHENIX. "Technical Design Report: sPHENIX experiment at RHIC". In: (2019). URL: https://indico.bnl.gov/event/5905/.

[5] Ferdinand Willeke. *Electron Ion Collider Conceptual Design Report 2021*. Feb. 2021. DOI: 10.2172/1765663. URL: https://www.osti.gov/biblio/1765663.